National Agency for the Evaluation of Universities and Research Institutes

anvur

Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca

Evaluation of Research Quality

vQr

Valutazione Qualità della Ricerca

Evaluation of Research Quality 2011-2014

(VQR 2011-2014)

ANVUR Final Report:
First Part: Statistics and summary results

21 February 2017

If you can't measure it, you can't improve it. Lord Kelvin

Not everything that can be counted counts, and not everything that counts can be counted. William B. Cameron, Informal Sociology: "A Casual Introduction to Sociological Thinking" (1963)

Table of contents

List of acronyms and specific terms

RESEARCH STAFF. The Institute's personnel who authored the research outputs under evaluation.

AM (Individuals in Mobility). Staff members who have been permanently recruited or have had a career promotion in the Institute in the VQR four-year period.

ANVUR. National Agency for the Evaluation of Universities and Research Institutes.

AREAS. The evaluation process has been divided into 16 scientific areas which are shown in the table below

CALL. The VQR 20112014 Call for participation.

BC. Cultural Heritage.

CETM Commission of Experts of the Third Mission evaluation. The Commission of Experts who evaluated the Third Mission.

CINECA. Inter-University Consortium for Computing. It managed the software interfaces and the administrative and accounting procedures during the evaluation.

CRC. Clinical Research Centres, specialized structures in clinical trials and assessed under TM, Health Protection.

CT. Third parties.

DM. The 27 June 2015 Ministerial Decree that appointed ANVUR to conduct the VQR 2011-2014.

ECM. Continuing Education Courses in Medicine, assessed under TM, Health Protection.

FC. Continuing Education.

GEV. Groups of experts for the Evaluation. The 16 panels of experts in the disciplines of the scientific areas that handled the evaluation of the research outputs submitted by the Institutes.

IRAS1-IRAS5. The research quality indicators by area and Institute are defined by the Call and are calculated as a fraction of an area's overall value.

IRFS. The final indicator of an Institute's research quality that integrates the area indicators IRAS1… IRAS5 with the weights attributed to the 16 areas.

IRD1-IRD3. The research quality by area and department defined by the Call, calculated as a fraction of the area's overall value.

IRDF. The final indicator of a department's research quality that integrates the IRD1-IRD3 indicators with the weights attributed to the 14 areas.

INSTITUTES. Institutes subject to the VQR evaluation. These are divided into: public and private Universities (which are obliged to be evaluated), MIUR-supervised Research Institutes (which are obliged to be evaluated), "similar" Research Institutes that requested evaluation under the same rules of the MIUR-supervised Research Institutes; Inter-University Consortia which requested evaluation using a subset of the indicators applied to Universities and MIUR-supervised Research Institutes; lastly, other Institute which requested evaluation under different but ANVUR-agreed rules.

LAW 240. Law no. 240 dated 30 December 2010 "Rules on the organisation of Universities, academic personnel and recruitment, and delegation to the Government to enhance quality and efficiency in the University system".

HANDBOOK. The document "Evaluating the Third Mission in Universities and Research Bodies. Evaluation Handbook" published by ANVUR in April 2015 in order to guide the evaluation of Third Mission data from the point of view of the criteria and evaluation questions.

MIUR. Ministry of Education, University and Research.

EXPECTED RESEARCH OUTPUTS The number of research outputs that each Institute was expected to submit for evaluation, obtained by multiplying each research staff member by the number of research output specified by the Call and adding up the results.

RESEARCH OUTPUTS Contributions defined in Section 2.3 of the Call (articles, monographs, book chapters, etc.) obtained by research and submitted for ANVUR evaluation.

SSD. The 370 Scientific-Disciplinary Sectors into which the 16 areas are organised.

SUB-GEV. GEV homogeneous subsets as defined by the scientific area's characteristics.

VQR. Evaluation of Research Quality

VQR1. Evaluation of Research Quality 2004-2010

VQR2. Evaluation of Research Quality 2011-2014

Foreword

One of ANVUR's tasks entrusted by the Presidential Decree 76 dated 1/2/2010 is the periodic evaluation of research and third mission activities of universities and research Institutes. Article 3 paragraph 1 letter a) reads: "*The agency ... assesses the quality of the processes, results and research outputs resulting from the management, training, and research activities, including technology transfer from universities and research Institutes ....*"

In July 2013, ANVUR completed its first Evaluation of Research Quality – VQR 2004-2010, which assessed research outputs and calculated other indicators for 2004-2010.[1] This report describes the activities and results of the second evaluation exercise, – VQR 2011-2014 which evaluated research outputs and calculated other indicators for 2011-2014.

The research results evaluation purposes are varied:

- it provides Italy with a fair and rigorous evaluation of research carried out in universities, research Institute and their internal structures (departments, Institutes,...), that can be used for different purposes;
  - ✓ it enables the Institutes' governing boards, to undertake actions to improve the research quality in the areas where they are weak compared nationally, or to enhance particularly promising or key areas for Italy;
  - ✓ it helps families and students make difficult choices related to study courses and universities;
  - ✓ young researchers can deepen their training and carry out research in the best departments;
  - ✓ Industries and public boards can address cooperation requests to the Institutes that host research groups, in terms of quality and critical mass, in the scientific areas which interest them;
  - ✓ *and much more…;*
- deciding a national ranking system by scientific area and Institute using the Call indicators to base the distribution of Universities' Fondo di Finanziamento Ordinario (Ordinary Financing Fund) reward share;

[1] The results can be seen on http://www.anvur.it/rapporto/.

- evaluating university departments and research Institutes' sub-Institutes for internal governing boards to autonomously guide the internal resource distribution;
- allowing a comparison of the national research quality against those of major industrialised countries.

The evaluation and results do not affect the quality of the teaching activities performed by universities. Its use helping young people approaching university becomes more appropriate where research plays a key role i.e. for master's and Ph.D. degree courses. ANVUR believes that good teaching in universities requires an adequate research activity at every level.

In addition, the rankings contained in the report, are the results of a research evaluation in the Institute which follows the Ministerial Decree (DM) and the Call, and should not be confused with the ranking of universities that some organisations, newspapers and universities publish annually. The rankings derive from broader assessments. They do not only cover research and its parameters involving universities of all countries but the far superior depth and detail of the research evaluation of Italian universities in the VQR. It is impossible to compare the rankings with the VQR results.

The VQR aims do not include a comparison of the research quality in different scientific areas. This is advised against by the parameters and different methods for evaluation of scientific communities within each area (for example, the prevalent use of bibliometrics in some areas and the peer review in others). It depends on factors such as the spread and national and international discipline references, different evaluation cultures or subjective ideas of what makes a scientific work "excellent" or "limited" in the various knowledge areas and the tendency between areas to unintentionally provide higher evaluations to improve their own discipline's position.

For viewing purposes the tables report the results of the evaluations in the various areas and should not be used to build merit rankings between the same areas. This needs different standardisation methods (as required by article 1, paragraph 319 of the budget law 2017).

Sometimes this caveat applies to the comparison between scientific-disciplinary sectors (SSD) which are internal to an area. In some cases, it is possible to compare the research quality between the SSD of the same area, in others (highlighted by individual area reports) such a comparison is impossible or undesirable. The area and homogeneous subsets rankings within an area, such as sub-GEV or SSDs, are aimed at internal national vertical comparison.

In building the Institute ratings (universities, research Institutes) it is necessary to provide a single indicator to rank the various areas where the Institute has carried out research. Throughout this report, it will be possible to see that the final Institute indicators are unaffected significantly by any differences in the assessment criteria used by individual areas.

The integration of the different area indicators into one Institute indicator similar to that used in the VQR1, requires the definition of the area weights, which can be done in different ways. The weights choice, and the methods for the use of the VQR results for the distribution of the Fondo di Finanziamento Ordinario share reward, are the Minister's responsibility.

The report contains a ranking of departments and sub-Institutes for each area. For the MIUR supervised research Institutes, the rankings for sub-Institutes are based on agreements reached between the Institutes and ANVUR.

It is important to take into account the associations choice of research output with the staff members which was made by the same Institute to optimise the Institute's overall assessment, putting less emphasis on departmental or sub-Institute evaluation. The departments and sub-Institutes' rankings included in this report provide information to the governing boards of Institutes which is to be used freely and independently, being aware of the above limits. The national and centralised research evaluation performed by ANVUR sets objectives and uses different methods from the "local" evaluation of departments conducted by individual Institutes. The two must co-exist, and the latter may enrich the other by providing contextual and programming elements that only local governing boards know and enhance. A local evaluation, carried out with faster and cheaper means can bridge the time gap between national evaluations, measuring progress and failures and preparing for timely interventions.

Finally, ANVUR emphasises that the **VQR results cannot and should not be used to evaluate individual researchers.** The reasons are many, and here we list the most important. The association choice of output with staff members is dictated by the Institute result optimisation and not the result of the individual subject. A request to publish only two research outputs in four years, in many science areas constitutes, a partial overall production of individual subjects; the non-consideration of the individual contribution to the research output with co-authors; and finally, evaluation methods where validity depends strongly on the size of the research group to which they were applied.

All the indicators described in the report are obtained as an average of elements belonging to heterogeneous populations – large generalist universities active in all or most areas with many researchers; medium and large specialised universities (such as the Polytechnics); small universities active in fewer areas; major research Institutes, like the CNR, which are active in all 16 areas; and traditional Institutes, present in many universities with affiliated researchers, but active in a single area, such as INFN and INAF. The indicators average values, starting from the Institute area assessment to that of sub-GEV, SSD and departments, have an increasing margin of statistical uncertainty, because the reliability of the average depends on the sample's size. However, when comparing the results of the two VQR (IRAS5 indicator) we have taken into account the uncertainty margin related to the classification of each Institute / area in each VQR.

ANVUR, for transparency reasons and because it will make the huge amount of data arising from the VQR2 available to the national and international scientific community, intends to make the VQR database public, after sensitive data is removed.

The report extension, the number and size of the tables and figures contained in the report have meant that the final report is structured into four parts. The first (this part) contains the text and appendices and includes comments upon the tables and figures. The second part is the detailed analysis of individual Institutes. The third is a comparison of the Italian research with the rest of the world, and the fourth, contains the analysis of the third mission activities in the Institutes.

All tables and figures of the first part, along with their caption are contained in the attachment in the order in which they are cited in the text. A second attachment, has the tables in Excel format to enable those who want to use analysis and sorting criteria other than those proposed in the text.

ANVUR and CINECA have managed and analysed the immense mass of VQR2 data with the utmost care. In compliance with the principles of reproducibility, ANVUR provides the original aggregated database at a sector/Institute level. Despite all the precautions taken and many cross-checks, some mistakes could appear in the difficult final coordination process. ANVUR can provide information and, where appropriate, correct any reported errors.

Sergio Benedetto

VQR 2011-2014 Coordinator

Daniele Checchi

ANVUR Governing Board

Marco Malgarini

ANVUR research area Executive

Rome, 21 February 2017

1 Introduction

The Evaluation of Research Quality 2004-2010 (VQR1) was one of the first activities which the ANVUR Governing Board focused immediately after its establishment, which took place on 2 May 2011. The first (and only) previous national research evaluation (VTR, Three-Year Research Evaluation) was conducted by CIVR for the years 2001-2003, with the output of the final report in February 2007.

Subsequently, the Ministerial Decree no. 8 dated 19 March 2010 laid down rules and procedures for the implementation of the second evaluation for 2004-2008, entrusted again to CIVR. The process stalled due to the output of the 1 February 2010 Presidential Decree 76 about the ANVUR establishment and operation, and the subsequent 22 February 2011 Presidential Decree which established the ANVUR Governing Board and appointed its members. Once established, ANVUR had to complete the operational programmes undertaken by CIVR, which ceased to exist and was replaced by the new agency.

The process was resumed with the 15 July 2011 Ministerial Decree, which replaced the previous 19 March 2010 Decree, and entrusted ANVUR with the execution of the Quality Research Evaluation for seven-years, from 2004 to 2010 (VQR1).

At the end of July 2011, ANVUR published a draft of the VQR1 Call on its website, and invited the universities and research Institutes to post comments, additions and proposals for amendments. The many suggestions received were reviewed and partly accepted in the final version of the VQR1 Call, which was approved by the ANVUR Governing Board in November 2011. With the output of the VQR1 Call on the agency's website, which took place on 7 November 2011, VQR1 officially started.

The evaluation was completed at the end of June 2013, (more than a month ahead of the schedule demanded by the Decree), with the output of the ANVUR Final Report and the 14 area reports (reducing the distance between the end of the observation period and data processing - up to two and a half years).

The second Evaluation of Research Quality (VQR2) was started with the output of the Ministerial Decree no. 458 dated 27 June 2015 (DM), which was followed by the output of the provisional call on the ANVUR site on 8 July 2015. Again, ANVUR asked to universities and research Institute to examine it by posting comments, additions and proposals for amendments.

The many suggestions received were reviewed and partly accepted in the final version of the VQR2 Call, (Call) published on the ANVUR site on 30 July 2015.

For the size and the limited time to devote to the preparation and conduct, VQR2, as VQR1, was a task of great complexity, which committed ANVUR substantial resources and the national scientific community. ANVUR expressed satisfaction in knowing that the beginning and conclusion of the VQR2 foreshadowed a periodic repetition of the evaluations, making it a structural and stable element of the national research system which takes place every five years, as stipulated in Art. 1, paragraph 339, of Law 232 dated 11 December 2016.

Many contributed to the completion of the VQR2, in several ways and degrees, and ANVUR thanks them for their cooperation. Those who particularly need to be mentioned are the outgoing and incoming ANVUR Board members, ANVUR directors, officers and employees, 16 GEV coordinators, with whom the VQR Coordinator worked hard and in great harmony, the 16 GEV assistants, who have experienced the VQR with great commitment and dedication, the 436 GEV members, who are valuable researchers who sacrificed their many commitments to the evaluation's success, the approximately 13,000 external reviewers who evaluated articles, monographs and the other research outputs carefully, and the Italian Publishers Association for its collaboration with ANVUR in successfully solving all the monographs' copyright issues by transferring encrypted files to CINECA. ANVUR thanks the CINECA working group coordinated by Pierluigi Bonetti, which, despite the other priorities which reduced the VQR2 commitment of some of the members, demonstrated a spirit of collaboration in responding to emerging needs.

Final thanks goes to the Institutes that participated in the VQR2. They operated in a spirit of great cooperation with ANVUR, in full awareness of the evaluation process' importance. ANVUR interpreted the Institute deadlines for the various VQR2 phases with flexibility and granted the requested extensions. It reopened the data transfer interface for the correction of the Call interpretation errors. This was due to the belief that the priority was obtaining reliable and comprehensive data for the development of indicators.

The "corpus" of research outputs resulting from the VQR2 is available on the ANVUR site in an easy to read format. It consists of the six parts of the final ANVUR report (four parts in

HTML text and pdf, charts and tables in pdf and excel format) and 18 area reports in pdf format.[2] The key VQR2 features and results are described below. The area reports, all approved unanimously by GEV, are a testimony to the spirit of collaboration and service that motivated them. They detail the progress and results of the evaluation in different scientific areas, deepen the sub-GEV and SSD area evaluation and contain many ideas to frame the assessment results of individual areas.

[2] A Third Mission / Impact activities report will be added to the 16 Area Reports.

2 The main VQR2 features

We suggest reading the Call to those interested in the regulation details and we summarise the evaluation's the main features in this section.

2.1 VQR2 participating Institutes

The evaluation covered compulsorily the universities and MIUR-supervised public research Institutes and allowed other Institutes conducting significant research activities to voluntarily undergo the evaluation and share the costs. All the organisations that participated will be identified by the generic term of "Institutes". Only type *a* and *b* permanent staff and researchers under Law 240 (staff members) took part in the VQR2. They presented two "research outputs"[3] published during 2011-2014 if they were university employees or three research outputs if they were research Institute employees or employees of the university with an official position at a research Institute.

The number of research outputs expected from each Institute was calculated considering the number of Institute staff members and/or those in charge of the research at the Institute and the number of research outputs that each staff member had to submit. This figure took into account the starting date of service for academic researchers and scientists and technologists of the research Institute and any periods of leave. The Call allowed reductions in the number of research outputs for staff members who had held Institutional positions (for details, see the Call).

96 universities took part in the VQR including 18 research Institutes where 12 MIUR-supervised and six similar research Institutes which had voluntarily asked to take part in the evaluation and to be compared with the supervised research Institutes. There were 21 other Institutes (nine inter-university consortia[4] and 12 research Institutes) who voluntarily asked to take part in the evaluation. The lists are shown in Tables 2.1 Table **2.2** and Table 2.3. Among the

[3] The term "output" refers to various kinds of contributions (articles, monographs, book chapters, etc.) which were published and based on research.

[4] The Semeion Consortium, which asked to participate in the VQR2 and had accredited staff members and submitted research outputs, decided not to share the evaluation costs. As a result, it will no longer appear in the tables in the following sections.

universities that were required to participate in the VQR2, IUL and Leonardo Da Vinci did not have credited staff members or sent context data, and therefore do not appear in the results.

Table 2.1. Universities participating in the VQR2

Table 2.2. MIUR-supervised research Institutes and volunteer research Institutes similar to a MIUR-supervised Institute participating in the VQR2

Table 2.3. Inter-university consortia and other Institutes participating in the VQR2 voluntarily

2.2 Research outputs submitted for evaluation

The types of research outputs accepted for evaluation were defined in the Call and further specified in the FAQ and subsequent News and evaluation criteria of the Groups of Experts for Evaluation (GEV).

Table 2.4 shows the distribution of the research outputs expected and submitted by universities and MIUR-supervised research Institutes in the VQR1 and VQR2.[5] Figure 2.1 displays percentages of the research outputs submitted by universities and MIUR-supervised research Institutes. Table 2.5 shows the distribution of the research outputs expected and submitted by all participating Institute to the VQR divided by area and type of output. The Institute provided the association of research outputs with the areas for the research outputs' evaluation. The table shows the number and percentage of monographs that staff members asked to be counted as two research outputs. You may notice that:

- the average percentage on the areas of missing research outputs is 5.9% (6.2% if we only consider the universities). This figure confirms an acceptable level of activity of teachers and researchers, and the attention of Institutes in meeting the Call requirements. The average percentage of missing research outputs in the VQR1 was 5.2% (4.7 for universities). The VQR2 figure is affected by voluntary abstention of some staff members, who, despite having published during the VQR2 period, have decided not to submit them for evaluation;

[5] The distributions of research outputs expected and submitted are less significant for voluntary Institutes, and therefore are not shown. This is due to the fact that there was no obligation for them to accredit all of their staff members.

- the missing research outputs' distribution in various areas is highly variable with percentages ranging from 2.8% to 10.1%; This variability is partly due to the distribution of inactive staff members (i.e. who have not submitted research outputs for evaluation) in the different areas. This reflects the destination of the research output determined by the Institutes, which is sometimes different from the staff member's area.
  In non-bibliometric areas, where the number of submitted monographs is significant, the percentage of monographs for which the staff member required that the evaluation counted twice was less than 10%. Except for areas 12 and 13, where the percentage was closer to 13%.

Table 2.4. Research outputs expected and submitted for universities and MIUR-supervised research Institutes in the two VQR evaluations

Figure 2.1. Percentage of the research outputs submitted by universities and MIUR-supervised research Institutes in the two VQR evaluations

Table 2.5. Expected and submitted research outputs by area and type. The area for each output for evaluation is indicated by the Institute

To accurately assess the missing research outputs percentages in the various areas Table 2.6 shows in the third column the research outputs submitted by staff members belonging to the area where the output is associated. Note how the variability in the distribution of missing research outputs decreases, with missing research outputs percentages ranging from a minimum of 3.1% to a maximum of 9.3%.

Table 2.6. Expected and submitted research outputs by area and type. The output area is that of the staff member to whom it was associated

As expected, Table 2.6 shows that for Areas 1-7, 8b, 9 and 11b, journal articles make up the majority of submitted research outputs, which are also the majority in the Area 13. In areas 10, 11a, 12 and 14, monographs and book contributions make up the majority of the research outputs. The bibliometric areas submit 94% of their scientific production in journal articles, while the same percentage drops to 43.2% in sectors which are not bibliometric (with oscillations ranging from 73% of the area 13 to 26% of the area 8.a).

Table 2.7 synthetically shows a comparison between the number of research outputs expected and submitted in the two cases of matching between the output and area of Table 2.5 and Table 2.6. The content of Table 2.7 is graphically displayed in Figure 2.2.

Table 2.8 shows the flow matrix of the research output submitted by member staff area (row) and output area (column). The matrix cell *(i, j)* shows the number of research outputs associated with the staff members in area *i* which were associated to evaluate area *j*. The main diagonal cells show the number of research outputs for which the staff members' area and the area indicated by the Institute for evaluation coincide.

Table 2.7. Summary of research outputs expected and submitted which emerge from Table 2.5 and Table 2.6

Figure 2.2. Histogram of the research outputs expected and submitted based on Table 2.7data

Table 2.8. Flow matrix of research outputs submitted by staff member area and assigned area of research outputs for evaluation

Table 2.9 and Table 2.10 show the same information in percentage referred to the rows or columns.

Table 2.9. Flow matrix of research outputs submitted by staff member area and assigned area of research outputs for evaluation in percentage related to the matrix rows

Table 2.10. Flow matrix of research outputs submitted by staff member area and assigned area of research outputs for a percentage evaluation related to the matrix columns.

The area that "conceded" the largest number of research outputs to other areas was area nine, while the area which received a greater number from the other areas was area six.

Table 2.11 and Figure 2.3 show the research outputs' distribution submitted for evaluation in different areas divided by output date during the VQR2 four-year period. The distribution over the years of output appears balanced, with the trend for some areas (in particular 12 and 13) to present a higher number of recent research outputs. The research outputs presented in years before 2011 and after 2014 in the table were included in the Call regulations concerning the date of output (electronic and/or paper format, for details see the Call).

Table 2.11. Distribution by area of research outputs submitted in the VQR2 four-year period

Figure 2.3. Histogram of the research outputs submitted by year (percentages of total research outputs between 2011 and 2014)

Table 2.12 shows the comparison between the percentage of various types of research outputs between the VQR1 and VQR2. An increase in the percentage of articles is noted. The number rose from 73.5% of the VQR1 to 78% of the VQR2. This increase is due to bibliometric and non-bibliometric areas. The monographs and book contributions decreased from 19.9% to 17.8%. Contributions in conference documents decreased from 5.8% to 3.3%.

Table 2.12. Comparison between the percentages of research output types in the two VQR

Table 2.13 and Table 2.14 show the numbers and percentages of the research outputs divided by area and language.

Table 2.13. Number of research outputs submitted by output area and language

Table 2.14. Percentage of research outputs submitted by area and language

Figure 2.4Figure 2.4 analyses the percentage distribution of the submitted research outputs by area and language. Overall, 76.6% of research outputs are in English. This percentage exceeds 90% in almost all "bibliometric" areas.[6] In non-bibliometric areas of human, legal and social sciences, Italian prevails. In area 10 – Classical, Philological-Literary and Historical-Artistic sciences 12.8% of submitted research outputs are non-English foreign language. This percentage drops to 6.1% in Area 11a.

Table 2.13. Number of research outputs submitted by output area and language

Table 2.14. Percentage of research outputs submitted by area and language

Figure 2.4. Histogram of the research outputs submitted by language

Table 2.15 shows the distribution of the number of authors for the output in absolute and percentage values, and Figure 2.5 shows the percentage graphically for each area. Table 2.16 shows a descriptive summary information of the distribution of the number of authors by output and Figure 2.6 shows the *pirate plot* of the same distribution, with an example which clarifies the reading.

Table 2.15. Distribution of authors for output in the 16 areas

Figure 2.5. Percentage distribution of authors for output in the 16 areas

Table 2.16. Descriptive information of the distribution of authors for output in the 16 areas

[6] In the report, the areas where most of the research outputs consisted of articles in indexed journals based on the Thomson Reuters and Elsevier B.V Scopus Web of Science databases, are called "bibliometric". These are areas 1, 2, 3, 4, 5, 6, 7, 8b (the Sub-Area of Engineering), 9 and 11b (the Sub-Area of Psychology). Area 13, which has similar behaviors (for internal sub-areas) to those of the bibliometric areas, can be placed between the two groups and use methods of output that refer to the styles of other neighbouring social sciences (such as in Area 14).

Figure 2.6. Pirate plot of the distribution of authors for output in the 16 areas

The tables and figures show different distributions of the number of authors by output, between the bibliometric and non-bibliometric areas, with an average number of authors by research outputs dropping from 208 in Area 2 to 1.1 in Area 12. In the bibliometric areas, it ranges from 208 in Area 2 to 3.4 in Area 1. Obviously, high numbers of authors by output can resubmit the same output, if permitted, by associating it with different staff members. For a detailed analysis of this phenomenon, see the Area Reports, including, in particular the Area 2 Report in which the phenomenon is particularly relevant.

2.3 Groups of experts for the Evaluation (GEV)

Unlike VQR1, the VQR2 DM has aggregated disciplinary research areas in 16 areas, each of which has appointed a Group of Experts for the Evaluation (GEV). The group number is in proportion to the number of research outputs expected in various areas with the aim of uniformly distribute the workload. Table 2.17 lists the 16 areas, the number of GEV and Coordinators' names. The table shows the number of the VQR2 GEV members who had already participated in the VQR1 under the same role. Their presence in the GEV has represented a positive element of continuity between the two evaluation periods. Small changes in the number of GEV members have been approved by the ANVUR Governing Board during the process, based on the number of research outputs delivered in the various areas. For the final number, any changes and the lists of GEV names, please refer to the area reports. ANVUR initially appointed 400 GEV[7] members, choosing a coordinator for each area.

Table 2.17. The 16 areas, the number of GEV members and coordinators

Appointment of GEV members was preceded by a rigorous selection process. This process initially focused on those who had answered the call published by ANVUR on 5 May 2015 indicating an intention to participate in the VQR2 evaluation.

The process used the following criteria:

[7] Due to the resignation of a limited number of GEV members, and the need to integrate some GEV composition the final number was 436. See the area reports for details.

1. the scientific quality (taking into account the scientific merit of the publisher, the number of citations, the research impact on the international community and any research prizes or other awards);
2. the continuity of the scientific research output in the last five years;
3. evaluation experience at a national and international level.

Selection of the candidates who passed the assessment according to the 1-3 criteria followed additional criteria:

a. coverage of cultural and research fields within the areas;
b. significant percentage of foreign university teachers;
c. fair gender distribution;
d. fair geographical distribution, where possible, for the candidates from Italian universities and research Institutes;
e. fair headquarters distribution, where possible, for the candidates from Italian universities and research Institutes.

In a limited number of cases, research extended outside the VQR2 candidate lists. This occurred when there were insufficient candidates with 1-3 characteristics for the area, or for cultural and research lines, or there were insufficient foreign university teachers or in cases where it was impossible to meet the a-e criteria.

Table 2.18 shows the percentages illustrating the correspondence of the lists for criteria a, b, c, shown previously. By comparing the percentage of women in GEV with the percentage of women among full professors (see Table 2.19) there is a greater presence of women in GEV.

Table 2.18. Distribution of GEV members

2.4 Staff members

Staff members were Researchers (permanent and temporary), Assistants, Associate Professors and Full Professors (permanent and temporary), under Article 12 paragraph 1 of Law no. 230 of 2005) of universities and Researchers, lead Researchers, Research Managers and Technologists, First Technologists and Technologist Director of MUIR-supervised research Institutes, in office on 1 November 2015. They were joined by staff members of Institutes which asked to participate in the VQR2 even if they were not obliged to do so.

The Technologists, First Technologists and Technologist Director of supervised research Institutes, who during the VQR period carried out exclusively administrative and service activities, were excluded from the evaluation.

Staff members belonging to the Institute where they were active on 1 November 2015, regardless of any previous affiliations, and research outputs associated with them were attributed to that Institute regardless of their affiliation at the time of output.

The universities' staff members, technologists, First Technologists and Technologist Director were required to deliver two research outputs, while researchers, lead researchers and research managers had to submit three. University Professors who had a formal research assignment (still active at the Call date) at a research Institute for at least two years (even if not consecutive) during the four-year period. The Call allowed for reductions in research outputs to be submitted for those who were recruited as university researchers and by research Institutes after 2005, or took periods of leave, or had held executive positions in their Institutes (see the Call for details).

Table 2.19 shows the distributions of staff members of universities and supervised research Institutes in the categories they belong, showing gender. The percentage of women among the staff members of the various areas varies from a minimum of 17.2% in Area 9 to 54.5% in Area 5, and is always lower than that of men with one exception. The overall percentage of women in the three main categories of universities and research Institutes, is modest, but significantly greater among researchers than among associate (or lead researchers) and full professors (or research managers). The gender distribution, taking into account the number of women graduates which was greater than male graduates, shows how difficult it is for women to access a researcher career. This was already shown in the VQR1.

Table 2.19. Distribution of staff members in various categories

2.5 Evaluation method

The research outputs evaluation delivered by the Institutes was made using the following methods either individually or in combination:

- direct evaluation by GEV, even using bibliometric analysis based on the number of research output citations and indicators of impact factors of the research output host journal;

- peer-review carried out by external and independent experts chosen by the GEV (usually two for each research output). Experts anonymously express their opinion on the quality of the research outputs to be evaluated.

The final quality opinion was based on the following criteria:

a) ***originality,*** *to be understood as the level at which the research output introduces a new way of thinking in relation to the scientific object of the research, and is thus distinguished from previous approaches to the same topic;*
b) ***methodological rigor,*** *to be understood as the level of clarity with which the research output presents the research goals and the state of the art in literature, adopts an appropriate methodology in respect to the object of research, and shows that the goal has been achieved;*
c) ***attested or potential impact*** *upon the international scientific community of reference, to be understood as the level at which the research output has exerted, or is likely to exert in the future, a theoretical and/or applied influence on such a community also on the basis of its respect of international standards of research quality.*

The evaluation result consisted of allocating the following merit classes to each output in terms of their weight:

- ***Excellent***: the research output was in the top 10% above the value scale shared by the international scientific community (weight 1);
- ***Good***: the research output was placed in the 10% - 30% segment (weight 0.7);
- ***Fair:*** the research output was placed in the 30% - 50% segment (weight 0.4);
- ***Acceptable***: the research output was placed in the 50% - 80% segment (weight 0.1);
- ***Limited***: the research output was placed in the 80% - 100% segment (weight 0);
- ***Ineligible for evaluation:*** the output belongs to types excluded from this exercise, or has attachments, and/or documentation which are inadequate for evaluation, or was published in the years before or after the four-year reference evaluation period (weight 0).

Each ***missing*** output compared against an expected number, was assigned a weight equal to 0.[8]

Each GEV approved its evaluation criteria, which ANVUR published between 17 and 20 November 2015.

ANVUR left a margin of autonomy to GEV interpretation and modulation of the criteria defined by the DM and the Call. Some elements are common to the various GEV. For more specific elements, each GEV chose a method which is more responsive to the regulations that compose it.

Elements common to all GEV:

- the ultimate GEV responsibility for the research outputs evaluation and the related allocation of merit classes;
- the choice to use the *informed peer review*[9], technique for evaluation, which consists of taking into account various assessment elements for the final merit classification. The elements range from the use of two databases for the bibliometric evaluation, to the combination of peer and bibliometric evaluation, depending on the GEV characteristics;
- the use of the *informed peer review* for evaluating monographs and book chapters;
- the procedure for the identification of external reviewers;
- the peer review performance includes a review sheet that contains three weighed multiple choice questions and the obligation to add a comment in support of the assessment;
- GEV operating rules;
- appropriate regulations to avoid conflicts of interest.

The common elements for all GEV (GEV01-07, GEV08b, GEV09, GEV11b) which could make use of the Web of Science (WoS) and Scopus databases for bibliometric evaluation are:

[8] This is an important innovation compared to the VQR1 which used to apply a penalty (with a weight equal to -0.5) for missing research outputs.

[9] *Informed peer review* refers to a review procedure that uses multiple sources of information to arrive at the final evaluation. For example, the final decision of a GEV internal consensus group can be based on the opinions of two external experts or bibliometric indicators.

- the use of two indicators, the first related to the citation impact of the journal that hosted the research output and the second is the citation numbers received by the single article;
- the cumulative distribution calculation of the two indicators in a disciplinary homogeneous category (such as an ISI WoS Subject Category) for the article's year of output evaluated using one of the two WoS and Scopus complete databases (i.e. not limited to the national records);
- the division provided by the two indicators in seven regions, of which five are for the allocation of one of five final classes, and two are characterised by contrasting information given by the two indicators, and thus require a peer review

GEV13 opted for a different evaluation algorithm, with a different weight between the bibliometric indicator (prevalent) and citation indicator (see Area 13 Report for further details on the subject).

GEV using bibliometrics adapted the assessment algorithm to their specific needs, while ensuring that they respected the percentage of research outputs in the various classes specified by the DM and the Call. For details please refer to the bibliometric GEV area Reports.

The common elements to all GEV (GEV08a, GEV10, GEV11a, the GEV12 and GEV14) that do not have sufficiently reliable databases and methods shared internationally for a bibliometric evaluation are:

- the generalised use of informed peer review to evaluate all research outputs.

3 Evaluation process

The evaluation of research outputs was carried out by GEV using bibliometrics and peer review. Each output was assigned two GEV members responsible for its evaluation process. The process will be described separately for each method below;

3.1 The *peer review evaluation*

The procedure related to the peer review evaluation began in February 2015 with the establishment of an ANVUR-VQR2 register of reviewers divided by GEV. Despite having the REPRISE reviewers register of MIUR available for the prior evaluation of the PRIN projects, it was considered proper to establish a new register. This considered that the REPRISE reviewers had never been subjected to prior assessment based on their scientific credentials, and that the number of foreign experts was limited.

GEV selected reviewers in the REPRISE register based on scientific merit (Hirsch index, number of citations, recent scientific research outputs) and, subsequently, have included a large number of experts selected using the same criteria and individually interviewed to assess their availability to participate in the VQR2. Obviously, the choice of merit criteria was modulated by the various GEV depending on the availability of bibliometric information.

For GEV12, an application form was released, to be used by those who were not included in the REPRISE Register, and wanted to contribute to the evaluation process as reviewers.

By integrating the REPRISE reviewers register lists with those prepared by GEV, the ANVUR-VQR2 list of nearly 14,000 names was created. The reviewers' selection process continued during the evaluation phase to involve expertise which had not been covered by the lists set up at that time, which became necessary for specific research output evaluation.

The two GEV members responsible for any output, separately chose two reviewers, to avoid conflicts of interest based on the information contained in the evaluation criteria documents.

Table 3.1, Table 3.2 Table 3.3 and Figure 3.1 show some statistics on the reviewers who participated in the VQR. They refer to Italian or "foreign" nationality. This refers to a connection to a foreign Institute and not the reviewer's nationality.

Table 3.1. Number of reviewers by area distinguished by nationality (Italian or foreign)

Table 3.2. Reviews assigned, made and rejected by area and nationality (Italian or foreign), with the exception of the reviews carried out internally by GEV members

Table 3.3. Number and percentage of total research outputs and research outputs subjected to peer review by area

Figure 3.1. Number of assigned reviews, carried out and rejected by area and nationality (Italian or foreign)

Overall, the VQR2 engaged 16,969 reviewers of which 13,546 were Italians and 3,423 had a foreign affiliation. The number of external reviewers (i.e. excluding the GEV members who acted as reviewers) identified as individuals is less than, and equal to 12,731, since the numbers in Table 3.1 added up the reviewers in each area, counting reviewers who were used in more than an SSD more than once. In Area 1 and Area 9 foreign reviewers are about 60% of the total, while in other areas Italian reviewers prevail. As seen in Table 3.2 and Figure 3.1 reviewers with an Italian affiliation were more available: 78% of the assigned research outputs were evaluated, against an equivalent value of 66% for foreign reviewers.

Table 3.3 shows that GEV 8a, 10, 12 and 14 assessed the totality of the research outputs using the peer method (in the table the percentages for such areas were slightly less than 100% because the research outputs relating to GEV members, but evaluated by other GEV, were also considered). It is important to emphasise that a sample, equal to approximately 10% of the research outputs evaluated bibliometrically, were also subjected to peer review to measure the degree of correlation of the two evaluation methods. A detailed analysis of the comparison method and its results can be found in Appendix A.

Part of peer reviews were carried out by GEV members, with the same external evaluation procedures. Overall, the percentage of peer reviews carried out directly inside GEV was small – 13.6%. Each research output subjected to the peer review was evaluated at least twice. In some cases, because of the delay in the delivery of the assessment by some reviewers, and the later submission to a third reviewer, the number of evaluations was greater than two.

Each reviewer evaluated the research output based on three multiple choice questions[10], one for each of the criteria a, b, c of Section 2.5. Each answer was assigned a score. The sum of the three scores was compared with four thresholds to generate a final classification into five classes.

[10] For questions and the scores, please refer to the Area Final Reports.

The classification enabled the reviewer to compare it with the definition of the classes 1, 2, 3 and 4 of Section 2.5 and, if necessary, to change the scores. The peer review required the formulation of a written opinion on the output, and the scores based on the three merit criteria.

Each GEV constituted consensus groups formed by two or three members and reached a final classification based on the scores expressed by two (or more) reviewers and a pre-defined procedure. The final assessments were approved firstly individually by the GEV Coordinator and then as a GEV consensus which could be carried out using telecommunication.

Despite small variations between various GEV, the procedure only needed the approval of the consensus group in cases of an identical peer review or a single class difference. If there were discordant evaluations for two or more classes a third peer review was asked for. Table 3.4 shows the absolute numbers and percentages of the research outputs that had conflicting reviews for one, two, three and four classes, for each GEV.

By mediating on all areas, the percentage of discordant reviews for at least two classes is equal to 19.7%. The topic is discussed in Appendix B, where we compare the bibliometric and peer reviews on a sample of research outputs for all GEV which could dispose of bibliometric indicators.

Table 3.4. Number and percentages of discordant peer reviews for 1, 2, 3 and 4 classes by area

3.2 Bibliometric evaluation

The bibliometric evaluation of GEV 2, 3, 4, 5, 6, 7, 8b, 9 and 11b covered articles published in journals indexed in WoS and Scopus databases. ANVUR acquired bibliometric information of archives for 2011-2014 for the world's scientific output from Thomson-Reuters and Elsevier, through CINECA. Unlike the choice made in other countries for similar evaluation exercises, ANVUR preferred to use both databases to avoid binding to a single manager, and to take advantage of the partial complementarity characteristics of the two databases.

Referring to the Area Reports for details on bibliometric algorithms used by each GEV, we briefly describe the main elements below.

The evaluation algorithm of GEV 2, 3, 4, 5, 6, 7, 8b, 9 and 11b is based on the calculation of two indicators for each output: the output citations and the impact of the journal. The Institutes were asked during the research output delivery, to indicate in the output sheet, the database (WoS or Scopus) and the impact indicator (IF5Y, Article Influence Score for WoS, and IPP SJR for

Scopus). For journals belonging to more than one Subject Category, Institutes expressed a preference which was submitted to the GEV for confirmation. Articles published by journals belonging solely to the multidisciplinary science category, which includes journals with a number of scientific subjects, such as Nature, Science, etc., have been assigned to another Subject Category based on (i) the citations contained in the article (ii) and the citations made to the article. For each of the journals one (or more) Subject Category was identified and the final one was selected based on a majority decision rule. When assigning a new Subject Category, the research output brought with it the impact factor of the publishing journal and the number of citations received, without changing the distribution destination of the Subject Category.

The pair of values of the output's characteristic indicators, with slightly different rules for each GEV (see the Area Final Reports), were associated with one of six classes: the five classes of the VQR2 and a sixth class (IR) obtained for divergent indicators (for example, a research output with a high number of citations published in a journal with very low impact or vice versa). The IR research outputs have undergone a peer review.

GEV1 adopted a slightly different bibliometric evaluation algorithm, which is not based directly on the ISI WoS and Scopus Subject Categories, but the reference categories, one for each SSD, for GEV. This integrates the Subject Categories (SC) used in WoS and All Science Journal Classification (ASJC) used in Scopus.   GEV1 used, in addition to the WoS and Scopus databases, and only for the indicator of the journal impact, the MathSciNet of the American Mathematical Society (MathSciNet). For details see the GEV1 Area Report.

GEV13 used a bibliometric algorithm significantly different from other bibliometric GEV, focusing on the publisher, and using the number of citations to "reward" the research outputs with a significant number of citations with a class jump. Again, for details please refer to the GEV13 Area Report.

While in the VQR1 self-citations were included without distinction from citations when calculating the citation indicator, in the VQR2 the number of self-citations exceeded the threshold of half of total citations. GEV members responsible for research output evaluation were asked to pay special attention to such cases.

Table 3.5 shows the absolute numbers and percentages of research outputs assessed bibliometrically and the IR research outputs for each GEV. The research outputs allocation to the areas is based on the staff member to whom they were associated. This explains why there are some research outputs in the areas 8a, 10, 11a, and 14 which were evaluated bibliometrically. For such research outputs, the Institutes suggested an evaluation of a bibliometric GEV different from the one to which the staff member belongs.

Table 3.5. Number and percentages of total research outputs and research outputs evaluated bibliometrically and IR classified by area

As already mentioned, the GEV algorithms for the bibliometric evaluation used different rules for the class allocation starting from the values of the two indicators. An accurate calibration of these algorithms to meet the percentages assigned to each class by the DM and the Call, was carried out before the criteria approval and output, allowing the Institutes to choose the research outputs under evaluation.

3.3 "Penalised" research outputs

The DM and the Call assigned zero score for "missing” research outputs, i.e. research outputs expected but not submitted by the Institutes, or those "ineligible for evaluation". Penalties for research outputs presented twice by the same staff member need defining. Table 3.3 and Table 3.5 shown above, list the number of missing research outputs and those ineligible for evaluation by area. The algorithm jointly decided by all GEV to attribute the penalties provides for five separate cases.

1. Each missing output receives a zero score.
2. The ineligible research outputs (the causes are varied and are indicated in the Call, as the lack of PDF files, or if the year of output is not included in the VQR2 four-year period, etc.) receive a zero score.
3. If an Institute has *n* times the same output, the latter is evaluated (e.g. with Excellent, score 1), while the others *n*-1 receive evaluation equal to zero. Each receives a score equal to 1/*n*.
4. If two different types of Institutes (for example, a university and a MIUR-supervised research Institute) associate the same research output with the same staff member, a research output is evaluated (e.g. excellent, score 1), while the other is penalised by a zero. Therefore, in each of the two a score of 0.5 is applied.

The lack of a pdf or an incomplete or illegible pdf did not automatically lead to a penalty; in such cases, ANVUR asked the Institutes to send the missing pdf (or replace the corrupted file). A penalty was only applied if the request was not met.

4 The indicators for the Institutes' research activity evaluation

The DM and the Call required an Institutes' ranking, and where possible their internal organisation (for example, departments), based on area indicators related to the research quality.

The report includes separate lists for universities and MIUR-supervised research Institutes; the other research organisations which chose to be evaluated using the same criteria; inter-university consortia and Institutes which agreed special evaluation rules with ANVUR.

Being aimed at the resource distribution, the Call indicators consider the quality expressed by the research outputs' evaluations, the information provided by the Institutes and their size. By using a combination of indicators with weights determined by the Call, each Institute was associated with a final indicator between zero and one. The sum of the indicator values of all Institutes belonging to a homogeneous group (universities, research Institutes, consortia, ...) is equal to one.

In the report the Institutes are also compared using three area indicators linked to the average research outputs' quality submitted regardless of the Institute size.

Below, we list the various indicators and illustrate the process that allows to pass from the area indicators to the Institute, department or sub-Institute final indicator.

4.1 VQR2 research activity indicators

The Call provided in the VQR2 five area indicators related to the research quality for the evaluation of universities and MIUR-supervised research Institutes, and other similar volunteer Institutes. For the evaluation of university departments or sub-Institutes of the research Institutes, the Call included five research quality indicators. For the reasons explained below, only three were calculated.

4.1.1 Area quality research indicators of universities, supervised and similar research Institutes

The area quality indicators included in the Call and used for universities, supervised and similar research Institutes which factor in the average quality and size of the Institute, are listed below with their related weights:

1. **The IRAS1 qualitative and quantitative indicator, (weight 0.75)**, is calculated as the ratio between the sum of the evaluations obtained by the research outputs submitted by the

Institute in a specific area and the overall evaluation of the area in homogeneous groups (universities, supervised and similar research Institutes, etc.).

2. **The IRAS2 qualitative and quantitative indicator, (weight 0.20)**, calculated as the earlier IRAS1 for a subset of evaluated research outputs and research outputs submitted by researchers who were recruited or promoted by the Institute in 2011-2014.
3. **The IRAS3 qualitative and quantitative indicator for resource attraction, (weight 0.01)**, is calculated by summing the funds obtained through participation in competitive Calls for national (PRIN, FIRB, FAR, ASI, PNR...) and international research projects (Framework Programmes of the European Union, European Space Agency, NIH, etc.). This value is expressed as a percentage of the overall value of the area in the homogeneous group.
4. **The IRAS4 higher educational qualitative and quantitative indicator IRAS4, (weight 0.01)**, is calculated as the number of PhD students, medical and health specialisation school students, research fellows, and post-doctoral personnel. This value is expressed as a percentage of the overall value of the area in the homogeneous group.
5. **The IRAS5 qualitative and quantitative improvement indicator, (weight 0.03).** Given the significant differences between the VQR 2004-2010 and VQR 2011-2014, the improvement indicator was not based on the values of the indicators obtained in the two evaluation exercises. Rather, it was based on the Institute's position in the distribution of a normalised version of the indicator *R* (defined below). The details of the algorithm for the calculation of IRAS5 are illustrated below.

All indicators described above, except for IRAS5 (also normalised), are expressed as a percentage of overall values of an area in the homogeneous group under evaluation. They depend on the "quality" and size of the Institute. If all the Institutes had the same average behaviour for the indicators, they would only reflect the size of the Institute in the specific evaluated area. The IRAS5 definition is more complex, and it refers to the subsection that describes the indicator.

4.1.2 Area quality research indicators for the inter-university consortia

The area quality indicators in the Call, which considered the average quality and the size of the inter-university consortia, are a subset of those used for universities and research Institutes which are based on consortia specific characteristics. These indicators are listed below with their related weights:

1. **The IRAC1 research quality indicator, weight 0.6,** equivalent to IRAS1
2. **The IRAC2 resource attraction indicator, weight 0.2,** equivalent to IRAS3
3. **The IRAC3 higher educational indicator, weight 0.1,** equivalent to IRAS4

4. **The IRAC4 improvement indicator, weight 0.1,** equivalent to IRAS5

Except for IRAC4, all indicators described above are expressed as a percentage of overall values of an area in the homogeneous group of consortia. They depend on the "quality" and size of the Institute. If all the Institutes had the same average behaviour for all indicators, they would only reflect the size of the Institute in the specific evaluated area.

4.1.3 Area research quality indicators of other volunteer Institutes

In addition to volunteer consortia and research Institutes which asked to be evaluated under the same rules used for the supervised research Institutes, other Institutes joined the VQR2 evaluation. They agreed the indicators and rules for the evaluation with ANVUR.

The area quality indicators in the Call, which considered the average quality and the size of the other volunteer Institutes, are a subset of those used for universities and research Institutes which are based on the volunteer Institutes' specific characteristics. These indicators are listed below with their related weights:

1. **The IRAE1 research quality indicator, weight 0.6,** equivalent to IRAS1
2. **The IRAE2 qualitative-quantitative indicator, weight 0.1,** equivalent to IRAS2
3. **The IRAE3 resource attraction indicator, weight 0.2,** equivalent to IRAS3
4. **The IRAE4 higher educational indicator, weight 0.1,** equivalent to IRAS4

All indicators described above are expressed as a percentage of overall values of an area in the homogeneous group of consortia. They depend on the "quality" and size of the Institute. If all the Institutes had the same average behaviour for all indicators, they would only reflect the size of the Institute in the specific evaluated area.

4.2 Quality indicators of the Institutes' scientific research outputs

The GEV were tasked to evaluate research outputs submitted by Institutes to collect information to compute the IRAS1, IRAS2, and IRAS5 indicators (in addition to IRAC1 and IRAE1 and IRAE2). In this section, we focus on the evaluation of the submitted research outputs quality, introducing indicators computed from the same information to be used to determine IRAS1.

Based on the VQR Call, individual research outputs were assigned weights equal to 1, 0.7, 0.4, 0.1 and 0 for Excellent, Good, Fair, Acceptable, or Limited, respectively. Missing research outputs were assigned a weight equal to 0;

Showing with $n_{i,j,EC}, n_{i,j,El}, n_{i,j,D}, n_{i,j,A}, n_{i,j,LIM}, n_{i,j,MAN}, n_{i,j,NV}$ the number of Excellent, Good, Fair, Acceptable, Limited, Missing, ineligible research outputs of the *i-th* Institute in the scientific-disciplinary *j-th* area , one obtains the overall evaluation $v_{i,j}$ of the *i-th* Institute in *j-th* area as:

$$v_{i,j} = n_{i,j,EC} + 0.7 \cdot n_{i,j,El} + 0.4 \cdot n_{i,j,D} + 0.1 \cdot n_{i,j,A} + 0 \cdot (n_{i,j,LIM} + n_{i,j,MAN} + n_{i,j,NV}) \quad (1)$$

In the next sections, we suggest three research quality indicators which are independent from the research staff numbers in the area's Institute. Then the IRAS1$_{i,j}$ indicator which takes into account research quality and the number of the research staff members evaluated in the Institute and belonging to the area.

The $v_{i,j}$ value is the basis for the calculation of the quality indicators for the research output we propose below.

Since Institute size is not taken into account, the first three indicators cannot be used by themselves for resource distribution. Nevertheless, they provide useful information on research quality for Institutes belonging to a certain area.

4.2.1 The first indicator

By indicating with $n_{i,j} = n_{i,j,EC} + n_{i,j,El} + n_{i,j,D} + n_{i,j,A} + n_{i,j,LIM} + n_{i,j,MAN} + n_{i,j,NV}$ the number of expected research outputs for the VQR2 of the *i-th* Institute in the *j-th* area, the **first indicator** $\boldsymbol{I_{i,j}}$, between 0 and 1, is given by:

$$I_{i,j} = \frac{v_{i,j}}{n_{i,j}} \quad (2)$$

It represents the average score obtained by the i Institute in the $j$area.

4.2.2 The second indicator

By indicating with $n_{i,j}$ the number of expected research outputs for the VQR2 the *i-th* Institute in the *j-th* area, and with N_{IST} the number of Institutes, the second indicator$\boldsymbol{R_{i,j}}$ is given by:

$$R_{i,j} = \frac{\frac{v_{i,j}}{n_{i,j}}}{\frac{\sum_{i=1}^{N_{\mathrm{IST}}} v_{i,j}}{\sum_{i=1}^{N_{\mathrm{IST}}} n_{i,j}}} = \frac{I_{i,j}}{V_j/N_j} \qquad (3)$$

where V_j and N_j indicate the overall assessment and the total number of expected research outputs in the *j-th* area within the homogeneous group of Institutes considered, namely:

$$V_j = \sum_{i=1}^{N_{\mathrm{IST}}} v_{i,j} \quad , \qquad N_j = \sum_{i=1}^{N_{\mathrm{IST}}} n_{i,j} \qquad (4)$$

The indicator $R_{i,j}$ is the ratio between the average score attributed to the expected research outputs of the i-th Institute in the j-th area and the average score awarded by all the research outputs of the j-th area. This allows a direct calculation of the research quality in a certain area for a particular Institute. Values of less than one indicate a scientific research output of a quality lower than the area average, values greater than one indicate a higher average quality.

4.2.3 The third indicator

The third $X_{i,j}$ indicator is the ratio between "excellent" and "good" research outputs of the Institute in the area and the ratio of excellent and good research outputs in the area within the set of homogeneous Institutes considered. Values greater than one of $X_{i,j}$ indicate that the Institute has a higher percentage of excellent and good research outputs than the area average. Formulas:

$$X_{i,j} = \frac{\frac{n_{i,j,EC+} + n_{i,j,EL}}{n_{i,j}}}{\frac{\sum_{i=1}^{N_{\mathrm{IST}}} (n_{i,j,EC+} + n_{i,j,EL})}{\sum_{i=1}^{N_{\mathrm{IST}}} n_{i,j}}}$$

4.2.4 The IRAS1$_{i,j}$ indicator, of the VQR Call

The $IRAS1_{i,j}$ indicator is defined in the VQR Call as the ratio between the total score achieved by an Institute in a given area and the total score of the area:

$$IRAS1_{i,j} = \frac{v_{i,j}}{\sum_{i=1}^{N_{\text{IST}}} v_{i,j}} = \frac{v_{i,j}}{V_j} \qquad (5)$$

It can be written as the product of the relative quality indicator of research outputs submitted by a specific Institute in an area and an indicator of the Institute's size in the same area. The quality indicator is a ratio between the average score received by the expected research outputs of the *i-th* Institute in the *j-th* area compared to the average score of all the expected research outputs in the *j-th* area, and corresponds to the first $R_{i,j}$ indicator defined in (3), while the weight of the Institute ($P_{i,j} = n_{i,j}/N_j$) is simply given by the share of expected research outputs in the *j-th* area due to the *i-th* Institute:

$$IRAS1_{i,j} = \frac{\frac{v_{i,j}}{n_{i,j}}}{\frac{\sum_{i=1}^{N_{IST,j}} v_{i,j}}{N_j}} \cdot \frac{n_{i,j}}{N_j} = \frac{I_{i,j}}{V_j/N_j} \cdot \frac{n_{i,j}}{N_j} = R_{i,j} \cdot P_{i,j} \qquad (6)$$

The $IRAS1_{i,j}$ indicator re-defines the weight of an Institute in an area, calculated by the share of expected research outputs, based on their relative quality. Thus, IRAS1 is a useful indicator for the allocation of funds across Institutes within the same area, since it considers both the Institute's quality and relative weight. The IRAS2, IRAC1, IRAE1 and IRAE2 indicators are defined in a similar manner.

4.2.5 The IRAS5 indicator

To calculate this indicator designed to measure the relative improvement between a VQR and the next, the Institutes were firstly divided into homogeneous groups (universities, research Institutes, consortia, etc.), which included the same Institutes in the old and new VQR. For universities, the group was further divided into three size classes, large, medium and small (for the thresholds that distinguish the classes, please refer to Table 6.1).

Then it will be the turn of the Institutes that were part of the new VQR and were excluded from the old.

To calculate the IRAS5 of the i Institute in the j area, an equivalence class is defined, so the universities are characterised by the reference indicator values that do not differ from each other in a statistically significant manner. The reference indicator is the standardised version of $R_{i,j}$.

The $\hat{R}_{i,j}$ standardised indicator is defined as:

$$\hat{R}_{i,j} = \frac{R_{i,j} - \mathrm{E}(R_{i,j})}{\sigma_j} \qquad (7)$$

where $R_{i,j}$ was defined in (3) and $\mathrm{E}\left(R_{i,j}\right)$ e σ_j respectively show the average and the standard deviation of the variable $R_{i,j}$ calculated on all area Institutes. For each *i* Institute we indicate by $\mathrm{N}_{\mathrm{P},\,i,\,j}$ the number of *k* Institutes with lower results such that

$$\hat{R}_{i,j} > \hat{R}_{i,k} + 1,\ \mathrm{k}=1,\ldots, N_{\mathrm{IST},j}$$

and with $N_{\mathrm{M},i,j}$ the number of *k* Institutes with higher results such that

$$\hat{R}_{i,j} < \hat{R}_{i,k} - 1,\ k=1,\ldots, N_{\mathrm{IST},j}$$

Once we define the variable

$$A_{i,j} = N_{\mathrm{P},i,j} - N_{\mathrm{M},i,j}$$

which represents the difference between the number of Institutes with an indicator statistically worse and the number of Institutes with indicator statistically better, each *i-th* Institute in the *j-th* area will be characterised by the two $A_{i,j}$ values calculated in reference to the old $(A_{i,j,\mathrm{V}})$ and new $(A_{i,j,\mathrm{N}})$ VQR.

To take into account the Institutes that are at the extremes of the distribution range we use specific criteria. We first consider the Institutes where Min $(A_{j,\mathrm{V}})$ $+3 < A_{i,j,\mathrm{V}} <$ Max $(A_{j,\mathrm{V}}) - 3$, and define the variable $\mathrm{B}_{i,\,j}$ as follows:

$$B_{i,j} = 0 \ \text{ if } \quad A_{i,j,\mathrm{N}} < A_{i,j,\mathrm{V}} - 2$$

$$B_{i,j} = 1 \ \text{ if } \ A_{i,j,V} - 2 \leq A_{i,j,\mathrm{N}} \leq A_{i,j,\mathrm{V}} + 2$$

$$B_{i,j} = 2 \ \text{ if } \quad A_{i,j,\mathrm{N}} > A_{i,j,\mathrm{V}} + 2$$

For the *i* Institutes such that $A_{i,j,\mathrm{V}} \geq$ Max $(A_{j,\mathrm{V}}) - 3$ the variable $\mathrm{B}_{i,\,j}$ is defined as follows:

$$B_{i,j} = 0 \ \text{ if } \quad A_{i,j,\mathrm{N}} < A_{i,j,\mathrm{V}} - 2$$

$$B_{i,j} = 1 \ \text{ if } \ A_{i,j,\mathrm{V}} - 2 \leq A_{i,j,\mathrm{N}} < A_{i,j,\mathrm{V}}$$

$$B_{i,j} = 2 \ \text{ if } \quad A_{i,j,\mathrm{N}} \geq A_{i,j,\mathrm{V}}$$

For the *i* Institutes such that $A_{i,j,\mathrm{V}} \leq$ Min $(A_{j,\mathrm{V}}) + 3$ the variable B$_{i,\,j}$ is defined as follows:

$$B_{i,j} = 0 \text{ if } \quad A_{i,j,\mathrm{N}} \leq A_{i,j,\mathrm{V}}$$

$$B_{i,j} = 1 \text{ if } A_{i,j,\mathrm{V}} < A_{i,j,\mathrm{N}} \leq A_{i,j,\mathrm{V}} + 2$$

$$B_{i,j} = 2 \text{ if } \quad A_{i,j,\mathrm{N}} > A_{i,j,\mathrm{V}} + 2$$

Finally, the Institutes which were excluded from the old VQR we defined the $B_{i,j}$ variable as follows:

$$B_{i,j} = 1 \text{ if } A_{i,j,\mathrm{N}} \text{ it is placed in the upper 50\% of the distribution}$$

$$B_{i,j} = 0 \text{ if } A_{i,j,\mathrm{N}} \text{ si situa nel 50\% inferiore della distribuzione}$$

The IRAS5$_{i,j}$ qualitative-quantitative indicator of the *i-th* Institute in the *j-th* area is obtained in the following manner:

$$\mathrm{IRAS5}_{i,j} \quad = \frac{B_{i,j} * n_{i,j}}{\sum_{i=1}^{N_{IST}} B_{i,j} * n_{i,j}}$$

Subsequently, the Institute IRAS5$_i$ indicator is obtained by adding the 16 area indicators multiplied by the area weights

$$\mathrm{IRAS5}_i \quad = \sum_{j=1}^{16} w_j \cdot \mathrm{IRAS5}_{i,j}$$

And finally, it is combined with other Institute IRAS indicators to obtain the final IRFS$_i$ indicator:

$$\mathrm{IRFS}_i \quad = \sum_{k=1}^{5} a_k \cdot \mathrm{IRASk}_i$$

where a_k are the weights of the indicators defined in the VQR Call.

4.2.6 The meaning of the area Institute indicators

The first $I_{i,j}$indicator between 0 and 1, represents the average score obtained by the *i* Institute in the *j* area.

The second $R_{i,j}$ indicator indicates the Institute's position compared to the area average within a set of homogeneous Institutes. If its value is greater than one, it means that the Institute has a quality above the average of the area, divided into homogeneous subsets by Institute type or size. If it is less than one the Institute is below average.

The third $X_{i,j}$ indicator provides information about the Institute's research outputs evaluated as excellent and good. If the value is greater than one, means that the Institute has achieved a percentage of research outputs evaluated as excellent and good, higher than the area average within the set of homogeneous Institutes.

Finally, the $IRAS1_{i,j}$ indicator, defined by the DM and the Call, combines a qualitative evaluation with the size of the Institute, and can be employed for a distribution of resources that can be viewed as a quality modification of a proportional distribution (based on staff members or on the number of expected research outputs). If in all Institutes all research outputs obtained the same average evaluation, then the indicator would reflect only the relative number of those submitted and the relative weight of the Institute within a specific area.

The area rankings of the Institutes presented in this Report and in 16 Area Reports were obtained using the $R_{i,j}$indicator.

4.3 Calculation of the Institute's final indicators

This section describes how to integrate the area indicator into the Institute's final indicator. The formulas and the text refer to the five indicators in the Call for universities and research Institutes. The application on inter-university consortia with a lower number of indicators, is obvious and omitted for brevity.

4.3.1 Institute qualitative and quantitative Indicator according to the Call

The five indicators listed in Section 4.1.1, all between zero and one with a sum equal to one in all the homogeneous Institutes (universities, research Institutes and consortia), are area indicators, they refer to the qualitative and quantitative positioning of an Institute in a specific area. The Institutes, however, carry out research in several scientific areas. To obtain an Institute order, it is necessary to include the area indicators in which the Institute carries out scientific activities with an overall Institute indicator that makes the end result less affected by assessment differences in different areas.

A solution to the problem of calculating the *i-th* Institute $IRFS_i$ final research indicator is as follows:

$$A_{i,j} = u_1 \cdot IRAS_{1,i,j} + u_2 \cdot IRAS_{2,i,j} + \cdots + u_5 \cdot IRAS_{5,i,j}\ ,\ j=1,\ldots,16 \quad (8)$$

$$IRFS_i = (w_1 \cdot A_{i,1} + w_2 \cdot A_{i,2} \ldots + w_{16} \cdot A_{i,16}\,) \qquad (9)$$

or, in summary:

$$IRFS_i = \sum_{j=1}^{16} w_j \cdot A_{i,j} = \sum_{j=1}^{16} w_j \cdot (\sum_{l=1}^{5} IRAS_{l,i,j} \cdot u_l) \qquad (9bis)$$

where:

- $IRAS_{1,i,j}$ is the IRAS1 indicator of the *i-th* Institute in the *j-th* area, similarly for $IRAS_{2,i,j}$ and so forth;
- $u_l\ , l = 1, \ldots, 5$ is the IRASl indicator weight (in brackets in the list 1-5 of Section 4.1.1), and
- w_j , *j* = 1, ..., 16 is the weight assigned to the *j-th* area.

The Institute IRFS final indicator is obtained by adding the five area and Institute indicators IRAS1, ..., IRAS5 in the Call, weighted with u_l weights assigned by the Call (formula 7), and then adding the $A_{i,\,j}$ Institute and area variables obtained, each weighted with the w_j area weight (formulas 8 and 9).

4.3.2 Weighting choice w_j

The weight definition of the w_j area is a MIUR "policy" choice. Choosing the w_j weighting is for guiding future research focusing on some areas over others, or impartially reporting the share of delivered research outputs or staff members of different areas or to be proportional to the share of financing historically assigned to the areas (for example, in PRIN and FIRB or European Calls).

5 Departmental research evaluation

The VQR had, among its tasks, to provide Institutes with a ranking of university departments in each respective area of research (or similar sub-Institutes in the case of research Institutes) that could be used independently as information from the Institutes' governing boards for internal resource distribution.

The universities' statutes approved after Law 240 created different types of departments. The most common were:

a. departments which bought together researchers belonging exclusively to an area;
b. departments that fully incorporated smaller pre-existing departments, with researchers that typically belong to one or two areas;
c. departments that include parts of pre-existing departments, with a strongly diversified structure which cannot be exclusively attributed to one (or two) areas.

In case a) the evaluation of the department area often coincides with that of the university to which it relates. In the other two cases, it was necessary to establish the department indicators starting from the research output evaluations associated with department's staff members which belong to different areas. It was important to ensure that differences of Inter-area evaluation did not significantly influence the result.

By indicating with $n_{i,j,k,EC}$, $n_{i,j,k,EL}$, $n_{i,j,k,D}$, $n_{i,j,k,A}$, $n_{i,j,k,LIM}$, $n_{i,j,k,MAN}$, $n_{i,j,k,NV}$, the number of Excellent, Good, Fair, Acceptable, Limited, Missing and Ineligible research outputs of the *k-th* department of the *i-th* Institute in the *j-th* area, we obtained the overall $v_{i,j,k}$ evaluation, of the *k-th* department of the *i-th* Institute in the *j-th* area as:

$$v_{i,j,k} = n_{i,j,k,EC} + 0.7 \cdot n_{i,j,k,EL} + 0.4 \cdot n_{i,j,k,D} + 0.1 \cdot n_{i,j,k,A} + 0 \cdot (n_{i,j,k,LIM} + n_{i,j,k,MAN} + n_{i,j,k,NV}) \tag{10}$$

5.1 Research quality indicators of departments and sub-Institutes in the area

Five quality indicators of the area were defined by the VQR Call. Based on the data provided by the Institutes about PhD students enrolled in specialised schools in the medical and health area, research assistants and post-doctoral fellows, which do not allow a precise allocation of these departments after law 240, the IRD4 indicator in the Call was not calculated. For reasons related to the different composition of many departments in the transition from the first to the second

VQR, the IRD5 improvement indicator in the Call was not calculated. The three area indicators IRD1, IRD2 and IRD3, took into account the average quality and the size of the departments and are listed below with their weights:

1. The research **quality indicator** (**IRD1, weight 0.75**), calculated as the sum of the scores obtained from the research outputs submitted. The value shall be expressed as a percentage of the overall area value.
2. **The qualitative and quantitative indicator IRD2, (weight 0.20)**, calculated as the previous IRD1 in the subset of research outputs and research outputs to be evaluated submitted by researchers who have been recruited or promoted by the Department in 2011-2014.
3. **The qualitative and quantitative indicator for resources attraction IRD3, (weight 0.05)**, calculated by summing the funds obtained through participation in competitive calls for national (PRIN, FIRB, FAR, ASI, PNR...) and international research projects (Framework Programs of the European Union, European Space Agency, NIH, etc.). The value shall be expressed as a percentage of the overall value of the Area.

The IRD1 indicator is calculated starting from the *R* area indicators, representing the average score of the department in the area divided by the area average score. The R indicator does not consider the differences of the score distributions among the examination sectors within the same area and is not standardised, i.e. it is not divided by the area index standard deviation. Because of the "standardised indicator of departmental performance" definition, required by Article 1, paragraph 319 of the 2017 Budget Law, ANVUR will deepen the appropriate homogeneous group for standardisation in the coming months, and develop a more appropriate standardisation method for the evaluation of the departments connected to teachers who belong to different areas and sectors.

5.2 Department’s scientific output quality indicators

As for the Institutes, in this section we introduce three quality indicators of the research outputs submitted by departments. These are independent of the staff member numbers being evaluated for the area within the departments. Since they do not take into account the department size, they cannot be employed by themselves to distribute resources, but need to be integrated with (or completely replaced by) the $IRD1_{i,j,k}$ indicator, which takes into account both research quality and the department size within the area. The first three indicators provide useful information about a department’s research quality within a scientific area.

5.2.1 The first indicator

By indicating with $n_{i,j,k}$ the number of expected VQR research outputs for the *k-th* department of the *i-th* Institute in the *j-th* area, the first $I_{i,j,k}$indicator, less than or equal to one, is given by:

$$I_{i,j,k} = \frac{v_{i,j,k}}{n_{i,j,k}}$$

and represents the average score obtained by the *k-th* department of the *i-th* Institute in the *j-th* area.

5.2.2 The second indicator

The second $R_{i,j,k}$ indicator is given by

$$R_{i,j,k} = \frac{\frac{v_{i,j,k}}{n_{i,j,k}}}{\frac{\sum_{i=1}^{N_{IST}} v_{i,j}}{N_j}} = \frac{I_{i,j,k}}{V_j / N_j} \qquad (13)$$

where V_j and N_j indicate the overall assessment and the total number of expected research outputs in the *j-th* area.

The $R_{i,j,k}$ represents the ratio between the average score received by research outputs of the *k-th* department of the *i-th* Institute in the *j-th* area and the average score received by research outputs in the *j-th* area. It allows a direct calculation of the relative research quality in a certain area, possibly divided into homogeneous subsets by Institute type or size, shown by a particular department. Values less than one indicate a scientific output with a quality lower than the area average, values greater than one indicate a quality which is higher than the area average.

5.2.3 The third indicator

The third $X_{i,j,k}$ indicator defined as the ratio between the fraction of excellent and good research outputs from the department in a specific area and the fraction of excellent and good research outputs of the area. Values greater than one of $X_{i,j,k}$ show that the Institute has a higher percentage of excellent and good research outputs than the area average.

5.2.4 The IRD1$_{i,j,k}$ indicator of the VQR Call

The *IRD*1$_{i,j,k}$ *indicator* is defined in the VQR Call as the ratio between the overall score achieved by a *k* department of the *i* Institute in a given *j* area compared to the overall evaluation of the area:

$$IRD1_{i,j,k} = \frac{v_{i,j,k}}{\sum_{i=1}^{N_{\text{IST}}} v_{i,j}} \qquad (14)$$

It can be written as the product between an indicator of relative quality of research outputs submitted by a specific department in a given area and an indicator of the department's size within the same area. The quality indicator is the ratio between the average score received by the *k-th* department research outputs of the *i-th* Institute in the *j-th* area and the average score received by all the research outputs in the *j-th* area, which corresponds to the third $R_{i,j,k}$ indicator defined in (13), while the size of the ($P_{i,j,k} = n_{i,j,k}/N_j$) department is simply given by the share of research outputs in the *j-th* area due to the *k-th* department of the *i-th* Institute:

$$IRD1_{i,j,k} = \frac{\frac{v_{i,j,k}}{n_{i,j,k}}}{\frac{\sum_{i=1}^{N_{IST}} v_{i,j}}{N_j}} \times \frac{n_{i,j,k}}{N_j} = R_{i,j,k} \times P_{i,j,k} \qquad (15)$$

The $IRD1_{i,j,k}$ indicator re-defines the weight of a specific department within a specific Institute in a given area. It is measured by the share of expected research outputs, based on their relative quality. As such, IRD1 is a useful indicator especially for the allocation of funds across departments within the same Institute in the same area, as it takes into account the research quality and the relative weight of the department.

The area rankings of departments presented in the16 area reports were obtained using the indicator $R_{i,j,k}$.

5.3 Departments and sub-Institutes ranking according to the indicators in the Call

The three indicators IRD1, ..., IRD3 (IRD2 and IRD3 were calculated for departments similar to the IRAS2 and IRAS3 indicators) described in Section 5.1. They were determined by data provided by the Institutes and the evaluation of research outputs. The indicator's final value was calculated for each department. This is linked to the *IRFD*$_{i,k}$ research, of the *k* department of the *i* Institute according to the following formula:

$$A_{i,j,k} = u_1 \cdot IRD1_{i,j,k} + u_2 \cdot IRD2_{i,j,k} + u_3 \cdot IRD3_{i,j,k}\ ,\ \ j = 1, \dots ,16,\ \ k = 1, \dots , N_{D,i} \quad \text{(16a)}$$

$$Q_{i,k} = w_1 \cdot A_{i,1,k} + w_2 \cdot A_{i,2,k} \dots + w_{16} \cdot A_{i,16,k} \quad \text{(16b)}$$

or, in summary:

$$Q_{i,k} = \sum_{j=1}^{16} w_j (\sum_{l=1}^{3} IRDl_{i,j,k} \times u_l) \quad (16c)$$

The final indicator is obtained by normalising the values $Q_{i,k}$ $s$according to the formula:

$$IRFD_{i,k} = \frac{Q_{i,k}}{\sum_{k=1}^{N_{D,i}} Q_{i,k}}\ ,\quad \sum_{k=1}^{N_{D,i}} IRFD_{i,k}\ = 1 \quad (16d)$$

where

- $IRD1_{i,j,k}$ is the IRD1 indicator of the *k-th* department of the *i-th* Institute in the *j-th* area, similarly for $IRD2_{i,j,k}$ and so forth;
- u_l , $l = 1, \dots ,3$ is the *IRDl* indicator weight (in brackets in the list 1-3 of Section 5.1.1), and
- w_j , *j* = 1, ..., 16 is the weight assigned to the *j-th* area.

The IRFD department final indicator is obtained by adding the three area, Institute and department indicators IRD1, ..., IRD3 weighed with u_l weights assigned by the Call (formula 16a), then adding the department, Institute and $A_{i,\, j,\, k}$ area variables obtained, each weighted with the w_j area weight (formula 16b), and, finally, by normalising the amount obtained by dividing by their sum on the Institute's departments (formula 16c).

The $IRFD_{i,k}$ indicator could be used directly to allocate department resources within the Institute in a way that takes into account the quality of the research department in the various areas and the department's staff members' numerical strength in the same areas. As detailed in the introduction, ANVUR's allocation to departments of the $IRFD_{i,k}$ indicator final value only provided guidance to the Institutes' governing boards, without intent to infringe their full autonomy in the internal method of resources distribution.

6 Research evaluation results for Institutes and departments

This section presents the results of the VQR related to the research quality. The tables' caption in the file is detailed to enable understanding even without reading the text.

In the first part, we will present a summary of the area evaluation results taken from the area reports. Subsequently, Institutes and departments will be compared within each area using only the evaluation of the research outputs based on the three quality indicators described in sections 4.2 (for Institutes) and the two indicators in Section 5.2 (for departments). Finally, the IRFS final Institute indicator described in Section 4.3 is calculated.

As mentioned above, the report shows separate ranking calculations for universities, research Institutes and inter-university consortia. For a better understanding of results, in each category of Institutes the tables and graphs separately show the large, medium and small Institutes, determined by thresholds on the number of expected research outputs depending on the area.

For the universities area ranking, the size thresholds for the 16 areas are indicated in Table 6.1. The thresholds were defined so that:

1. too different size classes were not used in both VQR (taking into account the different number of research outputs expected in both evaluation exercises);
2. class switching is realized for significant number differences in terms of expected research outputs; in other words, appropriately spacing the last university in a class from the first of the next;
3. take account of outliers in some areas (typically La Sapienza in Rome), which exhibit a high number of research outputs which excessively reduce the number of universities in the class G.

For area rankings of departments, the size thresholds were determined in the following manner:[11] the $n_{\mathrm{MAX},j}$ is the maximum number of research outputs expected for the departments in the j area. According to the DM, the results for the groups that include less than three staff members should not be published for reasons of insufficient statistical reliability and privacy protection. The

[11] In this report, the size thresholds for the departments have been calculated using the described algorithm. Some GEV (see the Area Reports) have made reasoned amendments to the size thresholds for departments.

big departments (G) in the *j* area are those with an expected number of research outputs included in the third higher range

$[5, n_{\text{MAX},j}]$, the average departments (M) are those with a number of expected research outputs in the third intermediate range $[5, n_{\text{MAX},j}]$. Small departments (P) are those with expected research outputs in the lower third range $[5, n_{\text{MAX},j}]$. The same principle guided the size distribution of SSD, sub-GEV and examination macro sectors rankings in the area reports. Due to the characteristics of these areas and depending on the departments, the size distribution of departments uses different thresholds for the three classes.

Table 6.1 shows the thresholds for expected research outputs in the universities in the 16 areas.

Table 6.1. Size class thresholds for universities in the 16 Areas

6.1 Areas research output evaluation results

In this section, we summarise the main results taken from the area reports. As mentioned in the report's introduction, the tables and graphics grouped for ease of reading the results of all areas, but note that it makes little sense to use them for a comparison between different areas.

Table 6.2 and Figure 6.1 show the overall numbers and the percentages of research outputs in the five VQR2 evaluation classes (A = Excellent, B = good etc.). The F column shows the amount and related percentage of missing research outputs and ineligible research outputs i.e. those research outputs which did not meet the Call evaluation criteria, because, for example, they were published outside the VQR2 four-year period, or because they were excluded from the GEV criteria.

Table 6.2. research outputs overall numbers and percentages in the VQR evaluation classes

Figure 6.1. research outputs overall numbers and percentages in the VQR evaluation classes

Table 6.3 and Figure 6.2 (only percentages) the overall numbers and percentages of research outputs in the VQR evaluation classes are divided by area. In the table, the column "A+B" also shows the sum of the research outputs belonging to the two "excellent" and "good" classes.

Table 6.3. Research output numbers and percentages by area in the VQR evaluation classes

Figure 6.2. Research output percentages by area in the VQR evaluation classes

Table 6.3 and Figure 6.2 show the distribution of research outputs in the VQR classes for all areas to allow the reader to find them in a single table. As mentioned in the Introduction, the table should not be used to compare the quality of scientific research outputs between different areas. The different values between the various areas of the percentages in the table depend on:

1. the percentage of research outputs evaluated with different methods (peer or bibliometric, see the comparison in Appendix B), which is different for each area;
2. the possible different "severity" of peer reviewers in the various areas;
3. the differences in the average quality of scientific research outputs.

In the impossibility of distinguishing the effect of point 3 from the first two on the distribution in the classes, any comparison between the different areas should be avoided.

6.2 The Institutes

Table 6.4. lists the universities alphabetically. For each university, the values of the three indicators of average quality of research $I_{i,j}$, $R_{i,j}$ and $X_{i,j}$ of Section 4.2 are shown, the two parameters *v* and *n* needed for their calculation, and their position in the ranking (overall and size class) for each area. The same information is contained in Table 6.5. for the research Institutes (supervised and volunteers to be compared with those supervised), in Table 6.4 for the other volunteer research Institutes that were not compared with the supervised in Table 6.7. For university consortia, the table shows the *R* area indicator which is calculated by considering the overall average of all participating Institutes to the VQR instead of just consortia.

Table 6.4. List of universities by area with the values of the average quality indicators of expected research output and ranking (overall and size class) for each area

Table 6.5. List of research Institutes (supervised and similar volunteers) with the values of the average research quality indicators and ranking for each area

Table 6.6. List of research Institutes (volunteers) with the values of the average research quality indicators and ranking for each area

Table 6.7. List of inter-university consortia with the values of the average research quality indicators and ranking for each area

Table 6.8. shows an evaluation summary of the universities, supervised and similar Institutes in the 16 areas. Each table row corresponds to an Institute (the Institutes are listed in alphabetical order within their respective types), and, for universities, the pairs of columns correspond to the 16 areas. The first column of each pair shows the type of Institute in the size class (Large, Medium, Small) and the second column shows the R indicator value of the Institute in the area. The cells' colour code has the following meaning: green is an Institute which occupies the first position in the area size class; blue is an Institute in the first quartile (but in a different position from the first) of the total distribution; red is an Institute in the last quartile of the overall distribution. An absence of colour indicates the presence of the Institute in the second or third quartile without distinction.

Table 6.8. Evaluation summary of universities and research Institutes in the 16 areas

6.3 Departments and sub-Institutes

Table 6.9. shows the universities in alphabetical order for each area. For each university, the departments to which staff members of that area belong are shown in alphabetical order. For each department, the values of the three average quality indicators of research $I_{i,j,k,}$, $R_{i,j,k}$ and $X_{i,j,k}$ of Section 5.2 are shown, the two parameters needed for their calculation, and the quartile to which they belong of a ranking built according to the $R_{i,j,k}$ indicator (overall and area size class). The calculation of the thresholds that distinguish the size class, was done according to the criterion described in the beginning of this section. These rankings are made by normalising the score of the research outputs submitted based on the average area score, and are comparable only within each area. To compare those departments that belong to different disciplines (or which cover several subjects), as required by Article 1, paragraph 319 of the 2017 Budget Law, you must determine the homogeneous group appropriate for the normalisation and the most appropriate method for standardising the departments' evaluation.

Table 6.9. List of university departments in alphabetical order with the values of the average quality research indicators and ranking (overall and size class) for each area

Table 6.10 contains the same information for the MIUR-supervised research Institutes which involve substructures in their internal organisation. This does not include the ranking position in the size class as the Institutes were not divided into such classes in this report.

Table 6.10. List of sub-Institutes of MIUR-supervised research Institutes with the average research quality indicators and ranking for each area

The indicators for the departments which have submitted less than five research outputs (for universities) and less than seven (for research Institutes) in a given area are not included due to insufficient statistical reliability or staff privacy.

6.4 Final Institute indicators

The five indicators IRAS1, IRAS2, ..., IRAS5 described in Section 4.1 were defined from data provided by the Institutes and the evaluation of research outputs. The final indicator value of an Institute was calculated and linked to the IRFS research defined in the formula (9) for each Institute. The ranking of Institutes was built separately for universities, research Institutes and interdepartmental consortia. Please note that the IRFS indicator uses five indicators of the VQR Call with their weights, and thus considers the Institutes' quality and size.

As can be seen from (9), the final indicator calculation requires the choice of w_j area weights. The values shown in the tables below used values such as the size values of the areas in terms of expected research outputs.

Table 6.11, Table 6.13, Table 6.15 and 6.17 show, for the Institutes (universities, supervised and similar research Institutes, consortia and other volunteer Institutes, respectively), listed alphabetically, the IRFS final indicator values (see formula (9) for universities, research Institutes, inter-university consortia and other volunteer Institutes. The context data required for the indicators calculation and the values of individual IRAS indicators in the Call are reported for each Institute in Part II of the report which analyses the individual Institutes in detail. Please note that the IRFS indicator values consider the size and quality of the Institute according to various parameters, and cannot be used to draw up a merit ranking.

The IRFS indicator values, which are added to one on the group of homogeneous Institutes, could be used directly as multiplication coefficients for resource distribution. The tables show the resource ratio coefficients that would be obtained using the relative weight of the Institutes measured by the number of expected research outputs. This allows checking the Institutes that would "benefit" from the VQR evaluation compared to a purely proportional distribution to staff members. The cells coloured in blue (red) have the higher (lower) IRFS values of the relative weight[12].

[12] In distributing the rewarding share of FFO in 2016, MIUR did not directly use the IRFS, but the values obtained without the IRAS5 indicator. See the related DM (http://attiministeriali.miur.it/anno-2016/dicembre/dm-29122016.aspx).

Table 6.12, Table 6.14, Table 6.16 and 6.18 show the weighted sum with area weights of IRAS indicators in the Call using the indicators calculated on the areas, for the Institutes (universities, supervised and similar research Institutes, consortia and other volunteer Institutes, respectively), listed in alphabetical order. The tables show the partition coefficients of the resources that would be obtained using only the relative weight of the Institutes measured by the ratio of expected research outputs, to allow the identification of indicators with a higher or lower relative weight for each Institute.

Table 6.11. List of universities in alphabetical order with the final IRFS Institute indicator values

Table 6.12. List of universities in alphabetical order with the values of the IRAS indicators in the Call weighted with area weighting

Table 6.13. List of research Institutes and similar volunteers in alphabetical order with the final IRFS Institute indicator values

Table 6.14. List of research Institutes and similar volunteers in alphabetical order with the values of the IRAS indicators in the Call with area weighting

Table 6.15. List of inter-university consortia in alphabetical order with the final IRFS Institute indicator values

Table 6.16. List of inter-university consortia in alphabetical order with the values of the IRAS indicators in the Call weighted with area weighting

Table 6.17. List of other volunteer Institutes in alphabetical order with the final IRFS Institute indicator values

Table 6.18. List of other volunteer Institutes in alphabetical order with the values of the IRAS indicators in the Call with area weighting

6.5 IRAS2 and IRAS5 indicators analysis

Besides IRAS1, two of the research indicators in the Call described in Section 4.1 (IRAS2, and IRAS5) depend on the evaluation of the research outputs submitted by the Institutes. In this section, we explain some summarised results for the two indicators. The IRAS2 indicator and recruiting quality in the Institutes.

The IRAS2 indicator (mobility indicator) is linked to the recruitment of the Institutes in the VQR2 four-year period. It is the ratio between the sum of the scores obtained by the permanently recruited staff members or have had a career advancement in the Institute and the total of area scores of staff members in mobility. As all the other indicators in the Call, IRAS2 consider the quality of the Institutes’ scientific research outputs and their number.

Three indicators were used to evaluate the Institutes’ recruitment policies (initially or for a higher position, staff in mobility, AM) compared to the scientific research output quality. The first, **area mobility R**, is the ratio of the AM average score of the Institute in an area and the average score of all area’s AM excluding the AM of the Institute under evaluation. If the ratio is greater than one, the Institute has hired or promoted on average staff members in the area with a VQR2 scientific research output better than the average of the AM in the area.

Table 6.19 shows the Institutes listed in alphabetical order for all areas within the two types of universities, supervised and similar research Institutes. The AM number, the first indicator value, the size class (Large, Medium, Small), the Institute's position in the overall and size class ranking (the latter only for universities) are shown for each area. The size class is defined according to the algorithm described in Section 6 for the departments. The green cells indicate that the Institute holds the first place in the area size ranking.

Table 6.19. Institutes listed in alphabetical order with mobility R indicator values of staff members in the Institute in the 16 Areas

Table 6.20 shows the Institutes listed in alphabetical order for all areas within the two types of universities, supervised and similar research Institutes. the AM number, the value of the second **R indicator referred to the area** which calculates the ratio between the average score of the Institute's AM in an area and the average score of the staff members in the area other than the area AM. This will highlight the trend that each Institute follows for recruitment positioning. The table for the universities shows the size class (Large, Medium, Small), the Institute's position in the overall and size class ranking. The size class is defined according to the algorithm described in Section 6 for the departments. The green cells indicate that the Institute holds the first place in the area size ranking.

Table 6.20. Institutes listed in alphabetical order where the R indicator values refers to the area of the staff members in mobility

Table 6.21 shows the Institutes listed in alphabetical order for all areas within the two types of universities, supervised and similar research Institutes. the AM number, the value of the third **R indicator referred to the Institute** which calculates the ratio between the average score of the Institute's AM in an area and the average score of the staff members of the Institute in the area other than the AM of the Institute in the area. This highlights any improvement that each area/Institute has done through recruitment. The *R* indicator describes the average score obtained in the VRQ by new hires/promoted in comparison to the permanent staff score. The table for the universities shows the size class (Large, Medium, Small), the Institute's position in the overall and size class ranking. The size class is defined according to the algorithm described in Section 6 for the departments. The green cells indicate that the Institute holds the first place in the area size ranking.

Table 6.21. Institutes listed in alphabetical order with the R indicator values of the staff members in mobility in the area

Table 6.20 and Table 6.21 highlight and a significant difference in the Institutes' recruitment policies, with very different ratio values.

6.5.1 The IRAS5 indicator and VQR1-VQR2 comparison

The IRAS5 indicator (improvement indicator) ranks differences of Institutes compared to the quality of the research outputs submitted for the VQR 2004-2010 and VQR 2011-2014. The weight of this indicator is low and amounts to 0.03 but represents the VQR's desire to highlight (and reward) Institutes that have shown tangible signs of improvement in some areas.

The definition of the IRAS5 indicator and the calculation method have been described in Section 4.2.5.

Table 6.22 gives a list of universities and research Institutes in alphabetical order with the values of the three $A_{i,j,\mathrm{V}}$, $A_{i,j,\mathrm{N}}$ and $B_{i,j}$ indicators (for their meaning see Section 4.2.5) for each of the 16 areas. In the table, the "Institute Positioning ... resulting from the distribution of R in the VQR1" column shows three cases of Institutes found in the central range, upper and lower extreme of the distribution of the R indicator standardised in the VQR1 (for details see the description of the algorithm in Section 4.2.5). Finally, the last column shows Institutes that were not included in the VQR1.

Table 6.22. Institutes listed in alphabetical order with the values of the $A_{i,j,\mathrm{V}}$, $A_{i,j,\mathrm{N}}$ and $B_{i,j}$ indicators in the 16 areas

Figure 6.3, displays a map of Italy with the main Italian universities marked by flags of three different colours showing the three $B_{i,j}$indicator values: green if it is worth 2, yellow if it is 1 and red if it is 0. Note that $B_{i,j}$= 2 indicates a marked improvement in the ranking position between the two VQR, $B_{i,j}$= 1indicates a stable position in the ranking, and $B_{i,j}$= 0 shows a deterioration in the ranking position.

Figure 6.3. Map of universities with colour codes for the $B_{i,j}$ indicator

6.6 Analysis of the scientific collaboration between Institutes

The Call allowed for the possibility that different Institutes would present the same research output if it was associated with different staff members, which is obviously only possible for research outputs with more than one author. The research outputs submitted by several Institutes are an indirect measure of the degree of collaboration between Institutes in different areas. It is significant to assess its quality by comparing it with the area average. Based on the research outputs submitted by several Institutes, we obtained the information in Table 6.23 and Figure 6.4.

The table shows the number of research outputs submitted by 2, 3, 4 and more than 4 Institutes, and the *R* indicator value for all categories in all areas. In this case, *R* is the ratio between the average score obtained from articles submitted by more Institutes and the average score of the area. As can be seen, and expected, the collaborations between Institutes are much more relevant in bibliometric areas, particularly in areas 2, 5 and 6. *R* is greater than one in all areas and for a number of participating Institutes which are more than one. The research outputs created from collaborations between several Institutes reflect prominent issues, potential leading of research outputs in accredited journals and many citations.

Figure 6.4 shows the distribution in the areas of percentages of excellent or good research outputs submitted by two or more Institutes.

Table 6.23. Distribution by area of the number and the *R* indicator of research outputs submitted by several Institutes

Figure 6.4. Distribution by area of percentages of excellent or good research outputs submitted by two or more Institutes

7 Conclusions

The VQR 2011-2014 has analysed a large amount of research data and evaluated more than 118,000 articles, monographs, and other research outputs published by Italian researchers of universities, MIUR-supervised research Institutes and other Institutes which asked to undergo the evaluation in the 2011-2014 four-year period. With the limitations mentioned in the report, the 16 area reports and the ANVUR final report, provide a complete overview of our country's quality of research for the Institutes (universities, MIUR-supervised research Institutes, volunteer research Institutes and inter-university consortia) and sub-Institutes that compose them.

Each GEV analysed the results of the evaluation in detail and published the results related to Institutes sub-Institutes for each of the sixteen areas and their subsets, up to the level of scientific-disciplinary sectors in the Area Report. The main goal of ANVUR's transparent publication of the results is to provide concrete evidence for all the stakeholders interested in the state of art of the Italian Research, in order to reflect and act upon, consolidate the strengths, repair the weaknesses and take corrective action.

We want to reiterate that the solution to problems can only start from an accurate understanding of the issues and the causes that generated them.

A complete analysis of the results, given their quantity, will require time and expert scientific work. To facilitate that task ANVUR intends to provide the evaluation basic data after any sensitive data has been deleted.

By comparing the results of Part IV of the report (international comparisons) with those of VQR2, we have an overview of a competitive Italian research scene compared to individual countries and country groups, despite Italy's retrogressive position in terms of researcher numbers and their financing.

The VQR2, albeit with a mitigation of the performance differences for hardly distinguishable reasons, such as the different classification of merit, the reduced number of research outputs, the bibliometric algorithm modified and improved and the positive effects of the evaluation culture, (as in the VQR1), prove that the good average quality of research is fairly heterogeneous. While there are universities that achieve positive results in many areas, there are universities which are below the area average Even with significant SSD and department level exceptions, this difference also shows a worrying gap between geographical areas, which may depend in part on contextual data that the VQR2 should not or could not analyse.

The evaluation process lasted 16 months and concluded in time for the result’s use in the distribution of the share reward for the FFO 2016. ANVUR, GEV, GEV coordinators’ assistants, the CINECA working group and the reviewers overcame many difficulties and, where necessary, corrected the procedure on the fly. In view of the next evaluation exercise, and to ensure a more successful outcome, we highlight some important aspects.

- The VQR tool is particularly suitable for evaluating homogeneous sets of significant sizes, such as universities. It reveals critical issues in the application of small strongly inhomogeneous sets such as MIUR-supervised research Institutes.
    - The evaluation should be extended to all research Institutes, regardless of the supervising Ministry. Otherwise the evaluation of research Institutes excludes key Institutes that might absorb a majority share of the funding (for example, research Institutes which depend on the Ministry of Health).
    - Of the eight MIUR-supervised research Institutes, two are not research Institutes in the strict sense of the word. The Italian Space Agency and the Consortium for Trieste's Scientific and Technological Area's main mission is developing and promoting scientific research, rather than carrying it out in-house. The remaining Institutes have different size and activity areas, CNR conducts research in all areas while others are limited to one or two areas.
- The decision to associate each research output to a staff member without allowing its reuse within the same university made the department’s evaluation weaker, because the selection of the research outputs aims at maximizing the Institute result.
- In the VQR1, research Institutes were forbidden to submit the same research output more than once by attributing it to different staff members belonging to sub-Institutes of the same Institute. In the VQR2 this was permitted for CNR, INFN and INAF. INFN could consistently reduce the total number of research outputs to be submitted and better select them (see the considerations of the GEV02 Report). It is presumable that in the future will be necessary to intervene on this aspect, for example with a higher limit to the number of times that the same research output is submitted.
- The peer reviewers’ selection process was thorough, and, as had already happened in the VQR1, it took into account the availability, scientific quality and expertise. One of the important results of the VQR was the setting up of a database of certified quality reviewers which is valuable to the agency's activities.
- The interface prepared by CINECA for accreditation of reviewers was the operation’s most serious bottleneck. It caused many delays and problems, and during summer 2016

the evaluation's conclusion was nearly postponed well beyond the end of the 2016. The VQR Coordinator suggested the creation of an archive of VQR reviewers which was independent from other CINECA-MIUR archives (such as REPRISE) and this would have solved the root problem but this met strong resistance inside CINECA. It was only in July 2016 that a last-minute agreement was reached which permitted the project to get back on schedule.

- The VQR2 analysed other important aspects related to research in addition to the evaluation of scientific research outputs. The recruitment quality analysis appeared statistically strong thanks to the relatively high numbers of newly hired and/or promoted staff members in the Institutes in the four-year period. Significantly, there is a strong correlation between the research evaluation research outputs results and the attention paid to recruiting the best researchers. It is a positive circle of cause and effect that makes us confident about the future improvement of the quality of research in our country.

- The identification of suitable indicators to assess the third mission activities is still an open issue. The term "third mission", unlike the first two missions (teaching and research) identifies these activities with an ordinal (third) instead of a defining noun and shows its provisional aspect. Compared to the VQR1, ANVUR set up a special commission of experts for the evaluation of third mission activities in the VQR2. The resulting analysis, described in detail in the second part of the ANVUR Final Report on the VQR2, is more accurate and stronger than that done in the VQR1. In VQR1, the third mission indicators only measured the amount of selected activities (patents, spin-offs, etc.), without analysing their specific characteristics or their quality. Despite the analysis improvements, however, ANVUR still considers the evaluation of third mission activities as experimental, and doubts that it is mature enough to be used for the purposes of resource distribution.

In conclusion, we believe that the VQR2 will unfold its beneficial effects in the months and years to come if its results are studied in detail, carefully analysed, and used by the Institutes' governing boards to start improvement actions. An encouraging sign is, once again, the evaluated Institutes' spirit of great interest and collaboration with ANVUR. The VQR2 required considerable work and commitment over a period which was far from easy, particularly for universities.